10.6 Conjugate Gradient Methods in Multidimensions

We consider now the case where you are able to calculate, at a given $N$-dimensional point $P$, not just the value of a function $f(P)$ but also the gradient (vector of first partial derivatives) $\nabla f(P)$.

A rough counting argument will show how advantageous it is to use the gradient information: Suppose that the function $f$ is roughly approximated as a quadratic form, as above in equation (10.5.1),

$$f(x) \approx c - b \cdot x + \frac{1}{2} x \cdot A \cdot x$$ (10.6.1)

Then the number of unknown parameters in $f$ is equal to the number of free parameters in $A$ and $b$, which is $\frac{1}{2}N(N + 1)$, which we see to be of order $N^2$. Changing any one of these parameters can move the location of the minimum. Therefore, we should not expect to be able to find the minimum until we have collected an equivalent information content, of order $N^2$ numbers.
In the direction set methods of §10.5, we collected the necessary information by making on the order of $N^2$ separate line minimizations, each requiring “a few” (but sometimes a big few!) function evaluations. Now, each evaluation of the gradient will bring us $N$ new components of information. If we use them wisely, we should need to make only of order $N$ separate line minimizations. That is in fact the case for the algorithms in this section and the next.

A factor of $N$ improvement in computational speed is not necessarily implied. As a rough estimate, we might imagine that the calculation of each component of the gradient takes about as long as evaluating the function itself. In that case there will be of order $N^2$ equivalent function evaluations both with and without gradient information. Even if the advantage is not of order $N$, however, it is nevertheless quite substantial: (i) Each calculated component of the gradient will typically save not just one function evaluation, but a number of them, equivalent to, say, a whole line minimization. (ii) There is often a high degree of redundancy in the formulas for the various components of a function’s gradient; when this is so, especially when there is also redundancy with the calculation of the function, then the calculation of the gradient may cost significantly less than $N$ function evaluations.

A common beginner’s error is to assume that any reasonable way of incorporating gradient information should be about as good as any other. This line of thought leads to the following not very good algorithm, the steepest descent method:

**Steepest Descent:** Start at a point $P_0$. As many times as needed, move from point $P_i$ to the point $P_{i+1}$ by minimizing along the line from $P_i$ in the direction of the local downhill gradient $-\nabla f(P_i)$.

The problem with the steepest descent method (which, incidentally, goes back to Cauchy), is similar to the problem that was shown in Figure 10.5.1. The method will perform many small steps in going down a long, narrow valley, even if the valley is a perfect quadratic form. You might have hoped that, say in two dimensions, your first step would take you to the valley floor, the second step directly down the long axis; but remember that the new gradient at the minimum point of any line minimization is perpendicular to the direction just traversed. Therefore, with the steepest descent method, you must make a right angle turn, which does not, in general, take you to the minimum. (See Figure 10.6.1.)

Just as in the discussion that led up to equation (10.5.5), we really want a way of proceeding not down the new gradient, but rather in a direction that is somehow constructed to be conjugate to the old gradient, and, insofar as possible, to all previous directions traversed. Methods that accomplish this construction are called **conjugate gradient** methods.

In §2.7 we discussed the conjugate gradient method as a technique for solving linear algebraic equations by minimizing a quadratic form. That formalism can also be applied to the problem of minimizing a function approximated by the quadratic form (10.6.1). Recall that, starting with an arbitrary initial vector $g_0$ and letting $h_0 = g_0$, the conjugate gradient method constructs two sequences of vectors from the recurrence

$$g_{i+1} = g_i - \lambda_i A \cdot h_i \quad h_{i+1} = g_{i+1} + \gamma_i h_i \quad i = 0, 1, 2, \ldots$$  \hspace{1cm} (10.6.2)
The vectors satisfy the orthogonality and conjugacy conditions
\[ g_i \cdot g_j = 0 \quad h_i \cdot A \cdot h_j = 0 \quad g_i \cdot h_j = 0 \quad j < i \] (10.6.3)

The scalars \( \lambda_i \) and \( \gamma_i \) are given by
\[ \lambda_i = \frac{g_i \cdot g_i}{h_i \cdot A \cdot h_i} = \frac{g_i \cdot h_i}{h_i \cdot A \cdot h_i} \] (10.6.4)
\[ \gamma_i = \frac{g_{i+1} \cdot g_{i+1}}{g_i \cdot g_i} \] (10.6.5)

Equations (10.6.2)–(10.6.5) are simply equations (2.7.32)–(2.7.35) for a symmetric \( A \) in a new notation. (A self-contained derivation of these results in the context of function minimization is given by Polak [1].)

Now suppose that we knew the Hessian matrix \( A \) in equation (10.6.1). Then we could use the construction (10.6.2) to find successively conjugate directions \( h_i \) along which to line-minimize. After \( N \) such, we would efficiently have arrived at the minimum of the quadratic form. But we don’t know \( A \).

Here is a remarkable theorem to save the day: Suppose we happen to have \( g_i = -\nabla f(P_i) \), for some point \( P_i \), where \( f \) is of the form (10.6.1). Suppose that we proceed from \( P_i \) along the direction \( h_i \) to the local minimum of \( f \) located at some point \( P_{i+1} \) and then set \( g_{i+1} = -\nabla f(P_{i+1}) \). Then, this \( g_{i+1} \) is the same vector as would have been constructed by equation (10.6.2). (And we have constructed it without knowledge of \( A \)!)  

Proof: By equation (10.5.3), \( g_i = -A \cdot P_i + b \), and
\[ g_{i+1} = -A \cdot (P_i + \lambda h_i) + b = g_i - \lambda A \cdot h_i \] (10.6.6)
with $\lambda$ chosen to take us to the line minimum. But at the line minimum $\mathbf{h}_i \cdot \nabla f = -\mathbf{h}_i \cdot \mathbf{g}_{i+1} = 0$. This latter condition is easily combined with (10.6.6) to solve for $\lambda$. The result is exactly the expression (10.6.4). But with this value of $\lambda$, (10.6.6) is the same as (10.6.2), q.e.d.

We have, then, the basis of an algorithm that requires neither knowledge of the Hessian matrix $\mathbf{A}$, nor even the storage necessary to store such a matrix. A sequence of directions $\mathbf{h}_i$ is constructed, using only line minimizations, evaluations of the gradient vector, and an auxiliary vector to store the latest in the sequence of $\mathbf{g}$’s.

The algorithm described so far is the original Fletcher-Reeves version of the conjugate gradient algorithm. Later, Polak and Ribiere introduced one tiny, but sometimes significant, change. They proposed using the form

$$
\gamma_i = \frac{(\mathbf{g}_{i+1} - \mathbf{g}_i) \cdot \mathbf{g}_{i+1}}{\mathbf{g}_i \cdot \mathbf{g}_i} \quad (10.6.7)
$$

instead of equation (10.6.5). “Wait,” you say, “aren’t they equal by the orthogonality conditions (10.6.3)?” They are equal for exact quadratic forms. In the real world, however, your function is not exactly a quadratic form. Arriving at the supposed minimum of the quadratic form, you may still need to proceed for another set of iterations. There is some evidence [2] that the Polak-Ribiere formula accomplishes the transition to further iterations more gracefully: When it runs out of steam, it tends to reset $\mathbf{h}$ to be down the local gradient, which is equivalent to beginning the conjugate-gradient procedure anew.

The following routine implements the Polak-Ribiere variant, which we recommend; but changing one program line, as shown, will give you Fletcher-Reeves. The routine presumes the existence of a function $\text{func}(\mathbf{p})$, where $\mathbf{p}(1:n)$ is a vector of length $n$, and also presumes the existence of a subroutine $\text{dfunc}(\mathbf{p}, \mathbf{df})$ that returns the vector gradient $\mathbf{df}(1:n)$ evaluated at the input point $\mathbf{p}$.

The routine calls $\text{linmin}$ to do the line minimizations. As already discussed, you may wish to use a modified version of $\text{linmin}$ that uses $\text{dbrent}$ instead of $\text{brent}$, i.e., that uses the gradient in doing the line minimizations. See note below.

```fortran
SUBROUTINE frprmn(p,n,ftol,iter,fret)
INTEGER iter,n,NMAX,ITMAX
REAL fret,ftol,p(n),EPS,func
EXTERNAL func
PARAMETER (NMAX=50,ITMAX=200,EPS=1.e-10)
C USES dfunc,func,linmin
Given a starting point $\mathbf{p}$ that is a vector of length $n$, Fletcher-Reeves-Polak-Ribiere minimization is performed on a function $\text{func}$, using its gradient as calculated by a routine $\text{dfunc}$. The convergence tolerance on the function value is input as $\text{ftol}$. Returned quantities are $\mathbf{p}$ (the location of the minimum), $\text{iter}$ (the number of iterations that were performed), and $\text{fret}$ (the minimum value of the function). The routine $\text{linmin}$ is called to perform line minimizations.
Parameters: $\text{NMAX}$ is the maximum anticipated value of $n$; $\text{ITMAX}$ is the maximum allowed number of iterations; $\text{EPS}$ is a small number to rectify special case of converging to exactly zero function value.
INTEGER its,j
REAL dgg,fp,gam,gg,g(NMAX),h(NMAX),xi(NMAX)
fp=func(p) Initialization.
call dfunc(p,xi)
do $j=1,n$
g(j)=xi(j)
h(j)=g(j)
```

Permission is granted for individuals to make one paper copy for their own personal use. Further reproduction or any copying for distribution is strictly prohibited. To order Numerical Recipes books directly, or for DRM-free PDF's, visit website http://www.nr.com or call 1-800-872-7423 (North America only), or send email to trade@cup.cam.ac.uk (outside North America).
Note on Line Minimization Using Derivatives

Kindly reread the last part of §10.5. We here want to do the same thing, but using derivative information in performing the line minimization.

Rather than reprint the whole routine linmin just to show one modified statement, let us just tell you what the change is: The statement

\[
\text{fret} = \text{brent}(ax, xx, bx, f1dim, tol, xmin)
\]

should be replaced by

\[
\text{fret} = \text{dbrent}(ax, xx, bx, f1dim, df1dim, tol, xmin)
\]

You must also include the following function, which is analogous to \texttt{f1dim} as discussed in §10.5. And remember, your function must be named \texttt{func}, and its gradient calculation must be named \texttt{dfunc}.

\begin{verbatim}
FUNCTION df1dim(x)
INTEGER NMAX
REAL df1dim,x
PARAMETER (NMAX=50)
USES dfunc
INTEGER j,ncom
REAL df(NMAX),pcom(NMAX),xicom(NMAX),xt(NMAX)
COMMON /f1com/ pcom,xicom,ncom
do 11 j=1,ncom
   xt(j)=pcom(j)+x*xicom(j)
endo 11
call dfunc(xt,df)
df1dim=0.
END
\end{verbatim}
do :: j=1,ncom
   df1dim=df1dim+df(j)*xicom(j)
endo::
return
END

CITED REFERENCES AND FURTHER READING:

10.7 Variable Metric Methods in Multidimensions

The goal of variable metric methods, which are sometimes called quasi-Newton methods, is not different from the goal of conjugate gradient methods: to accumulate information from successive line minimizations so that $N$ such line minimizations lead to the exact minimum of a quadratic form in $N$ dimensions. In that case, the method will also be quadratically convergent for more general smooth functions.

Both variable metric and conjugate gradient methods require that you are able to compute your function’s gradient, or first partial derivatives, at arbitrary points. The variable metric approach differs from the conjugate gradient in the way that it stores and updates the information that is accumulated. Instead of requiring intermediate storage on the order of $N$, the number of dimensions, it requires a matrix of size $N \times N$. Generally, for any moderate $N$, this is an entirely trivial disadvantage.

On the other hand, there is not, as far as we know, any overwhelming advantage that the variable metric methods hold over the conjugate gradient techniques, except perhaps a historical one. Developed somewhat earlier, and more widely propagated, the variable metric methods have by now developed a wider constituency of satisfied users. Likewise, some fancier implementations of variable metric methods (going beyond the scope of this book, see below) have been developed to a greater level of sophistication on issues like the minimization of roundoff error, handling of special conditions, and so on. We tend to use variable metric rather than conjugate gradient, but we have no reason to urge this habit on you.

Variable metric methods come in two main flavors. One is the Davidon-Fletcher-Powell (DFP) algorithm (sometimes referred to as simply Fletcher-Powell). The other goes by the name Broyden-Fletcher-Goldfarb-Shanno (BFGS). The BFGS and DFP schemes differ only in details of their roundoff error, convergence tolerances, and similar “dirty” issues which are outside of our scope [1,2]. However, it has become generally recognized that, empirically, the BFGS scheme is superior in these details. We will implement BFGS in this section.

As before, we imagine that our arbitrary function $f(x)$ can be locally approximated by the quadratic form of equation (10.6.1). We don’t, however, have any